EPIDEMIOLOGY & BIOSTATISTICS SUMMARY

SideKick

YOUR BEST FRIEND

Introduction to Epidemiology

I- Definition of Epidemiology: *"upon people"*

It's the study of the **frequency**, **distribution** & **determinants** of diseases and health –related states or events in **specific population**, and the application of this study to the **prevention and control** if diseases and other health problems.

- **Major purpose of epidemiology** → obtain, interpret & use health information to promote health and to eliminate or reduce health problems.

Study

How?
=> Through:
- observation
- surveillance
- hypothesis-testing research projects
- analysis of epidemiologic data

Distribution

* Time: مرض معين بيحصل في الصيف/الشتا
* Place: rural/urban
* Persons: males/females

= Descriptive epidemiology

Determinants

ايه المسئول عن المرض؟
- biological, chemical, physical, social or behavioral factors.

= Analytic epidemiology

II- Uses of Epidemiology:

1. In Health-care management:

a- assessment of community health needs (community diagnosis) to set policy & plan health programs and to determine whether health services are available, accessible, effective and efficient.

Through answering the following questions:

1. What are potential health problems?
2. Where are they?
3. Who is at risk?
4. Which ones are declining?
5. Which ones are increasing?

} answered by methods of descriptive and analytic epidemiology

b- Planning, implementation & evaluation of services and setting priorities.

c- Developing health policies.

2. In understanding the disease process:

a- searching for causes or risk factors of diseases.

b- study natural history (&prognosis) of diseases.

c- historic study of rise & fall to determine factors utilized in prevention or causing re-emergence of disease.

d- establish a clinical diagnosis of disease.

3. In public health practice:

a- investigations of epidemics.

b- surveillance of diseases.

c- making projects.

d- assessing programs used for mass screening of diseases.

e- assisting in formulating medical teaching curricula.

4. In clinical and preventive practice:

a- assessing effectiveness of preventive and treatment modalities.

b- helps in defining and classifying diseases.

c- studying prognostic factors.

d- studying effectiveness of diagnostic modalities.

e- assessing in clinical decision making.

f- basic science for planning, conducting, & analyzing clinical research.

III- Sources of Epidemiology:

Source	Importance
1. Population statistics **a) population census**	For planning of social and health services *"census is enumeration at specific point of time of individuals comprising population within the area → regarding: age – sex – economic date"* - used as denominator for: 1. morbidity rates 2. moratlity rates 3. fertility rates يعني لما آجي أحسب الحاجات دي بعمل معادلة حسابية وهي مثلا عدد الناس المريضة بمرض معين على عدد السكان الكلي فتطلعلي النسبة
b) population estimates	- لما بنحتاج معلومات محدثة ولسه الcensus فاضل عليه وقت على ما يتعمل فبيتم حساب عدد السكان تقديريًا ساعتها.
2. Vital records **a) Birth registration**	For legal, demographic and epidemiologic purposes. - They provide the denominator for: 1- Infant mortality & morbidity rates 2- Maternal morbidity & mortality rates & numerator for: 1- fertility rates لما تقسم عدد المواليد على عدد السكان الكلي = fertility rate
b) Death registration	- Basic documentation for: 1- determining the number of deaths. 2- calculating mortality rates & ratios. 3- determining causes of deaths.

Source	Sterngths	Weaknesses
3. Notification of infectious diseases: **a) International** **b) National reporting**	1- Forward incidence data. 2- Indicate fluctuation of occurrence of diseases. 3- Predict occurrence of epidemics. 4- Allow tracing source of infections. 5- provide data to control.	1- The number of cases reported is much lower than reality. 2- Not all health problems are notifiable. يعني في حالات إصابة بالمرض هتعدي من غير ما تتسجل أو يتبلغ عنها ويكون عدد المرضى الحقيقي أكتر من اللي اتسجلوا.
4. Disease registers e.g. TB – Cancer – DM – IHD – Mental disorders	1- good source for accurate information on: - incidence-prevalence rates - duration of diseases - cure, mortality & disability rates 2- To study environmental hazards	1- Expensive 2- Need active research program 3- Not all diseases are registered.
5. Hospital records	1- correct diagnosis 2- To study serious and severe conditions 3- To study hospital acquired diseases	1- Information may not be complete 2- not representative of the population 3- usually deal with severe conditions
6. Special subgroup records	Diagnosis is correct	Not representative to the population
7. Morbidity surveys		
8. Record linkage هنا بنربط بين معلومات من أكتر من مصدر عشان نطلع تسلسل للتاريخ المرضي للشخص مثلاً	1- Provide full course of illness of the individual 2- provide information on: a- natural history of specific diseases/morbidity b- Family/genetic disorders c- chronic history of specific disease or morbidity d- use of health services	

Measures of disease frequency

"Morbidity & Mortality rates"

Tools of measurement:

- → **Counting the number of cases.**
- → **Ratio:** $\frac{a}{b}$ "numerator isn't included in denominator"
- → **Proportion:** $\frac{a}{b}$% "numerator is included in denominator"
- → **Rate:** $\frac{a}{b}$ "at a given period of time" *(Rate is like time but at a given time)*

Utilities of counting	Limitations of counting	Uses of ratios, proportions & rates
1. Readily available data 2. To monitor occurrence of important infectious diseases 3. For administrative purposes: a- indicate load of illness b- to allocate resources	Can't be used to compare disease load. بمعنى اني لو قلت اسكندرية فيها 3 حالات فيروس سي والقاهرة فيها برضو 3 حالات، كده انا مقارنتش كويس لأن ممكن في اسكندرية يكونوا 3 حالات من اصل 100 فرد وفي القاهرة من اصل 1000 فرد فكده يكون المرض منتشر في اسكندرية اكتر :3	1. Describe morbidity, mortality and fertility. 2. Characterize population by age, sex, ethnic background.

Morbidity measures

	Incidence rate	Prevalence rate	
Formula	$\frac{\#\text{New cases}}{\#\text{people @ risk}} \times 10^x$ *"in a given time frame"* * *we usually use census population counts*	$\frac{\#\text{Current cases (old \& new)}}{\#\text{ total n. of people}} \times 10^x$ * *It depends on:* 1. Incidence rate 2. duration of disease	
Factors affecting it	1. ↑↑ people at risk → ↑↑ incidence rate 2. changes of the etiological factor → ↑↑ incidence rate 3. quality of preventive program → ↑↑ quality → ↓↓ incidence rate 4. improvement of diagnostic tools → ↑↑ incidence rate "apparently"	**Increased by**	**Decreased by**
		1. ↑↑ new cases 2. longer duration of disease 3. prolongation of life without cure "↑↑care" 4. In-migration of cases "↑↑numerator" 5. out-migration of healthy people "↓↓ denominator" 6. improvement of diagnostic methods	1. ↓↓ new cases 2. shorter duration of disease 3. ↑↑ case-fatality rate "↑↑ mortality" 4. In-migration of healthy people "↑↑denominator" 5. out-migration of cases "↓↓numerator" 6. improvement of cure rate of cases

Utilities	1. planning of health services 2. Indicator for evaluation of preventive programs 3. To measure absolute risk from total sample 4. used to calculate: - relative risk - attributable risk - attributable risk percent "Cohort study"	1. planning of health services

* **Attack rate:** special form of Incidence rate applied to a narrowly defined population.
"فئة معينة من الناس تجمع بينهم حاجة واحدة،، مثلاً:

- outbreak in a school
- puerperal sepsis
- people attended a party & developed gastroenteritis

- **- Formula:** $\frac{\#\text{new cases}}{\text{people @risk}} \text{x}10^{n}$ *"for a limited time"*

* **Secondary Attack rate:** a measure of frequency of new cases among contacts of the patient

- Formula: $\frac{\#\text{secondary cases among contacts}}{\text{total \#contacts}} \text{x}10^{n}$

- Utilities: to measure ease to communicability of diseases

Note: when you measure the total number of people at risk or the total number of contacts, you should exclude the immune persons who were vaccinated or previously ill.

* **Point prevalence rate:** a snapshot look at the population "at a particular time"
* **Period prevalence rate:** a look at the population over a longer period "at a range of time"

* **Recovery rate:** $\frac{\#\text{recoveries}}{\text{total \#patients}} \text{x}10^{n}$

Mortality measures

*** Mortality rate = Crude death rate:** $\frac{\#\text{deaths}}{\#\text{midyear population}} \times 10^n$ *"at a particular time"*

*** Age-specific death rate:**

1. **Infant mortality rate:** $\frac{\#\text{deaths under 1yr}}{\#\text{live births}} \times 1000$ => *Indicates social & health conditions of the country*
2. **Neonatal mortality rate:** $\frac{\#\text{deaths under 28 days}}{\#\text{live births}} \times 1000$ => *Indicates maternity care*
3. **Post-neonatal mortality rate:** $\frac{\#\text{deaths 28 days to under 1yr}}{\#\text{live births}} \times 1000$

*** Sex-specific death rate:** $\frac{\#\text{deaths of a specific sex}}{\#\text{population of the same sex}} \times 10^n$ *"at a particular time"*

*** Maternal mortality rate:** $\frac{\#\text{deaths due to pregnancy,delivery,pueperum \& associated conditions}}{\#\text{live births}} \times 100.000$

"at a particular time"

=> *Indicates status of maternal health & services*

*** Cause-specific death rate:** $\frac{\#\text{deaths due to specific cause}}{\#\text{midyear population}} \times 10^n$ *"at a particular time"*

*** Proportionate mortality rate:** $\frac{\#\text{deaths due to specific cause}}{\#\text{deaths due to all causes}} \times 100$ *"at a particular time"*

*** Case-fatality rate:** $\frac{\#\text{deaths due to specific cause "disease"}}{\#\text{patients having same disease}} \times 100$ *"at a particular time"*

=> *measures pathogenicity & virulence of the cause of the disease*

Descriptive Epidemiology

➔ According to definition of epidemiology:
"The study of distribution (=descriptive epidemiology) and determinants (=analytic epidemiology) of health related status….

➔ Descriptive epidemiology is concerned with: **Person-Place-Time** triads which describe health related phenomena.

➔ Each item will be discussed in different dimensions to get a precise description of the state, So what are the uses of descriptive epidemiology?
1. It helps in formation of etiological hypothesis.
2. It provides data for planning & evaluation of preventive and curative services.

I- Person Distribution:

1. **Age:** interpretation of age distribution of disease is explained according to:

a- stage of development:
- congenital anomalies can only result from intra-uterine exposure to dangerous influences.
- children lack immunity to infections than adults.
- sexually-transmitted diseases are more common among sexually active adolescents & young adults.
- fractures in old people due to osteoporosis.

b- degree of exposure, susceptibility & duration of immunity:
- chicken pox is a disease of childhood.

c- hormonal changes:
- Breast cancer shows two peaks: ➔ **40-49 years** due to ovarian dysfunction
  ➔ **after 65 years** due to adrenal estrogen imbalance

d- cumulative effects:
- Chronic diseases e.g. atherosclerosis tend to increase with age.

Importance of studying age distribution:
1. **Severity of diseases:**
   - pneumococcal infections are more severe in children & elderly.
   - osteoporosis → severe in elderly & bone cancer → severe in young.
2. **Clinical type of disease:**
   - cretinism → children
   - myxedema → adults
3. **Bimodality**: if the disease has 2 peaks in different ages, this suggests 2 mechanisms of the disease.

2. Sex:

- **Some diseases occur more frequently in males as:**
  lung cancer – coronary heart disease
- **Some diseases occur more frequently in females as:**
  Diabetes – obesity – hyperthyroidism

➔ **Differences in sex-related diseases are due to:**

a- Anatomical differences:
- Prostate cancer → males … Uterus cancer → females

b- Sex-linked genetic inheritance:
- Hemophilia "non-dominant X-linked disease"

c- Hormonal factors:
- Coronary heart disease → common in young men & pot-menopausal females

d- Differences in habits & behavior:
- smoking in men → chronic bronchitis

e- Variations in infectious diseases:
- Tetanus → common in men

3. Marital status: married people have lower mortality rates than singles, due to:

- people with dangerous life tend to be single
- people with diseases tend to remain single

→ Suicide & mental illness are more common in singles
→ Breast cancer is more common in singles
→ Cervix cancer more common in those who marry young

- Couples usually develop same diseases as they have similar dietetic, psychological, social and environmental influences.

4. Occupation: It is an important epidemiological factor because:

1. It's associated with social class
2. Exposure to certain working conditions/risks → diseases
3. It may alter the general pattern of life

5. Socio-economic status: It is measured by:

1. Educational level
2. Occupation
3. Income
4. Residence
5. Housing condition

*** Diseases which are more common in low socio-economic classes:**

1. TB
2. Rheumatic heart disease
3. chronic bronchitis
4. malnutrition
5. Infant, maternal & preschool mortalities

*** Diseases which are more common in high socio-economic classes:**

1. Obesity
2. Gout
3. Coronary heart disease
4. atherosclerosis

6. **Behavior:** smoking – eating habits – sedentary life – drug abuse …
7. **Religion:**
   1. Alcoholic liver cirrhosis & Taenia solium don't appear in Muslims
   2. Cervix cancer is less in Muslims and Jewish due to personal hygiene – male circumcision
8. **Racial/Ethnic factors:**
   1. ↑↑ Hypertension frequency in black
   2. ↓↓ TB in European Jews

   } Due to:
   1. Hereditary factors.
   2. Social & environmental influences

II- Place Distribution:

General characteristics peculiar to place:

1. all ethnic groups in one area have the same frequency of the disease
2. same ethnic groups outside the area don't show high frequency of the disease
3. healthy persons entering the area become ill
4. inhabitants leaving the area don't show disease

Variations in place distribution of disease:

1. International variations:

- Breast cancer rates are highest in western countries & lowest in Japan

2. National variations:

- Endemic goiter "due to iodine deficiency" is more frequent in deserts

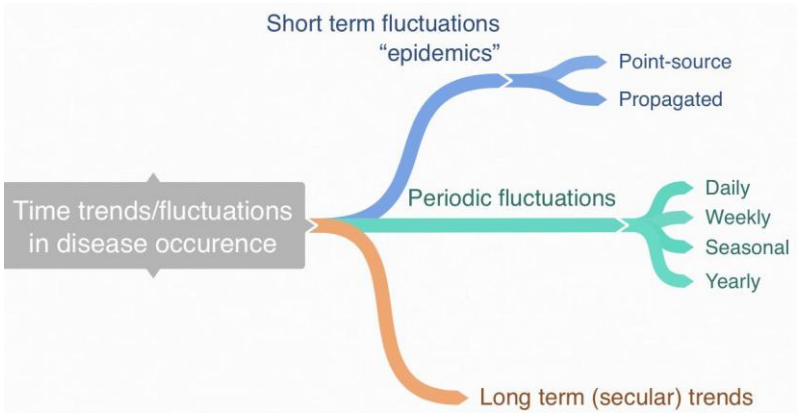
3. Rural/Urban variations:

- Chronic bronchitis, accidents, lung cancer, cardiovascular diseases, mental illness & drug dependence → more common in urban areas
- Zoonotic diseases, bilharziasis → more common in rural areas
- Fertility, death rates → high in rural areas

4. Local variations:

- TB, scabies, rheumatic fever → more common in slummy areas

III- Time Distribution:



1. Short-term fluctuations "Epidemics":

Epidemic: The occurrence of cases of illness in a community or region, in excess of normal expectancy.

Types of epidemics:

a- Point-source epidemics:

1. Simultaneous exposure to the disease agent
2. Resultant cases are explosive in number
3. all cases develop in short period of time "one incubation period"
4. Epidemic curve rises & falls rapidly "no secondary waves"

as: **Cholera**

b- Propagated epidemics:

1. Infections are propagated by passage from person to person.
2. Cases occur over a longer period of time "more than one incubation period"
3. Epidemic curve rises & falls gradually

as: **Measles**

2. Periodic fluctuations = cyclic trends:

a- Daily variations:

Hay fever *"allergic condition"* → diurnal rhythm, due to: Housekeeping activities "sweeping" & going to field

b- Weekly variations:

Car accidents → weekends, due to: Reckless driving – alcohol – drug consumotion during weekends.

c- Seasonal variations:

- Upper respiratory infections → winter
- GIT infections → summer

d- Yearly variations:

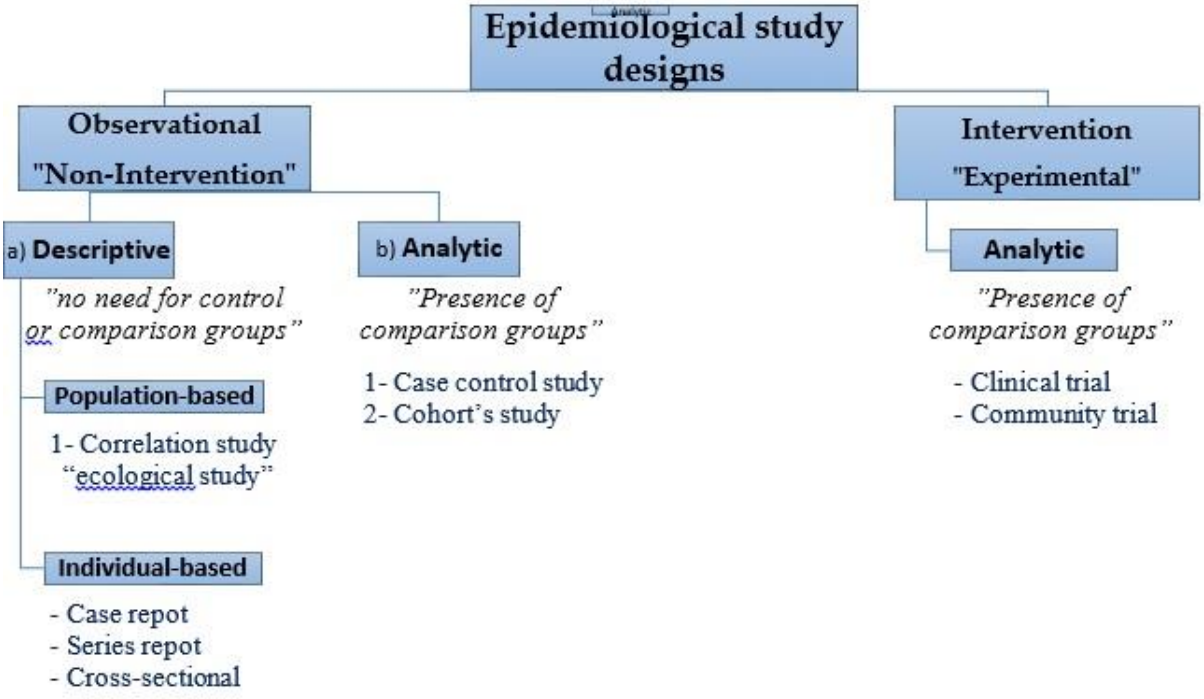
Measles in pre-vaccination era appeared in cycles every 2-3 years

3. Long-term fluctuations = secular trends:

* Several years or decades are needed to see changes in occurrence of disease "increase/decrease":

- Coronary heart disease – lung cancer - diabetes → have upward trend during the past 50 yrs.
- Puerperal sepsis – diphtheria – polio → have downward trend

Epidemiological study designs



I- Observational Studies:

a- Descriptive studies:

They are important for:

1- Providing information on the frequency of health states and their known and possible causes.
2- Guiding planning of health promotion and disease prevention activities.
3- Providing important clues as to the causes of different health states.
4- Providing a base for analytic/experimental studies.

a.1. Correlation "ecological" study:

It occurs at the level of an entire population measuring the association between 2 quantitative variables.

Examples:

- Correlation between cigarette's sales and mortality from lung cancer. "Positive correlation" *as both variables are increasing or decreasing together.*
- Correlation between percentage of decline in mortality from cervix cancer and percentage increase in number of women screened for cervical cancer. "Negative correlation" *one variable is increasing while the other is decreasing.*

Correlation coefficient (r): → **It is a value that lies between -1 and +1, so;**

-1 means: strong negative correlation "relationship"
+1 means: strong positive correlation "relationship"
0 means: no correlation "relationship"

Advantages of correlation studies:

1. Easy, quick, cheap
2. Use information already available
3. useful in formulating hypothesis

Disadvantages of correlation studies:

1. Inability to link a particular exposure "risk factor" to a particular individual
   "because it is made at the level of population"
2. Inability to control for **confounding factors.**

Confounding factor: a factor that affects both Exposure and Disease, so distorting the relationship between them & confound (confuse) the results.

e.g. if we study the association between smoking & lung cancer, the different ages of people of people at the study will confuse the results, *"as lung cancer increases with age"* hence age is called a confounding factor.

---

a.2. case report and case series:

case report describes a new unusual phenomenon in one case

case series describes a new unusual phenomenon collected from individual case reports.
"more than one individual with the same phenomenon"

Advantages:

1. Easy, cheap
2. Useful in formulating hypothesis
3. Individuals in case series may serve as cases in a case control study *"see later"*

Disadvantages:

1. Based on the experience of a single individual, it might be a coincidence
2. problem of finding a suitable comparison group *"see later"*

a.3. Cross-sectional study "prevalence study":

Exposure (E) and Disease (D) states are assessed simultaneously among individuals in a well-defined population.

There will be 4 possibilities:

Disease / Exposure	+ve	-ve
Yes	a	b
No	c	d
Total	a + c	b + d

a: have the disease & exposure "risk factor"
b: don't have the disease but have the exposure
c: have the disease but don't have the exposure
d: don't have the disease and the exposure

From the previous data, you can measure:

Prevalence rate of the disease among the exposed individuals: $\frac{a}{a+b} x100$

Prevalence rate of the disease among non-exposed individuals: $\frac{c}{c+d} x100$

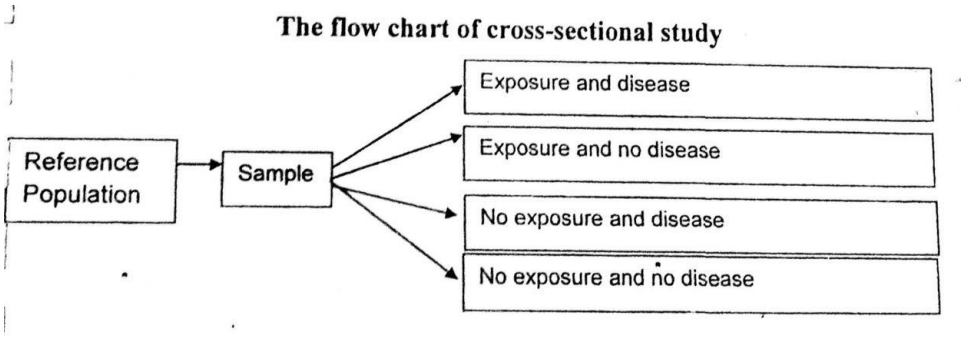
Prevalence rate of the disease in this population: $\frac{a+c}{a+b+c+d} x100$

Advantages of cross-sectional study:

1. Easy, quick, cheap
2. Directs case finding → يعني بيتم اكتشاف المرضى اللي مكانوش عارفين مرضهم
3. Generates hypothesis
4. Estimates prevalence rate in relation to exposure
5. shows case load → *as it measures prevalence rate*
6. Reveals distribution if health and disease states
7. provides information for planning and evaluation of health services

Disadvantages of cross-sectional study:

1. Not useful in rare diseases → لأنك احتمال كبير مش هتلاقيهم عشوائي كده وسط الناس
2. Not useful in acute diseases → لأنه بيخلص بسرعة فمش هتلاقيه برضو
3. Deals only with survivors (survival bias) → بما إنك بتعامل بشكل عشوائي ففي الغالب العيان مش هتلاقيه أو يكون مات أصلا
4. Fails to demonstrate the temporal relationship → إيه اللي حصل الأول الdisease ولا الexposure؟



b. Analytic Observational studies:

b.1. case-control study:

It's an observational (non-intervention) analytic study in which individuals are selected on the basis of having the disease "cases" or not having the disease "controls" and then compared in relation to the presence or absence of a particular exposure (risk factor).

- يعني أنا بقارن بين مجموعتين واحدة عندها المرض والتانية سليمة من حيث مدى تأثير exposure معين على كونهم مرضى أو أصحاء

Steps of case-control study:

1. **Selection of cases:**
   a- set selection criteria
   b- gather cases from sources (hospitals – general population)
   * we find selection bias in hospitals → as it reflects the characteristics of a group of patients attending a specific facility.

2. Selection of controls:

a- set selection criteria

b- gather individuals from sources (hospitals – general population – neighbors – friends …)

c- Matching:

we select the controls in way they are similar to the cases regarding certain confounding factors "that influence outcome of the disease"

أنا هنا عايز يبقى عندي factor واحد بس هو اللى بيأثر على ظهور المرض من عدمه

d- Size of the controls:

- one control for each one case
- may be more than one (2-3) for each one case and this increases the precision of the study
- 4 controls → increase the cost with little precision

3. Determine the exposure under study using the same methodology for both cases and controls:

a- review of records → يعني هتبدأ تراجع سجلات المستشفى وتشوف حالات المرضى بتوعك وكده

b- Interview → مقابلة شخصية معاهم أو استبيان عشان تعرف طبيعة مرضهم وكده

4. Analysis and interpretation of results:

a- Tabulation of data

Disease / Exposure	+ve	-ve
Yes	a	b
No	c	d
Total	a + c	b + d

b- calculation of exposure rates among cases and controls:

* دايماً الكلمة اللي بعد among بتكون هي الdenominator

- Exposure rate among cases = $\frac{a}{a+c}\ x100$
- Exposure rate among controls = $\frac{b}{b+d}\ x100$

c- Estimation of the disease risk associated with exposure "odds ratio"

OR

* Measure of the strength of association between the risk factor and disease.

→ odds ratio = $\frac{a\ x\ d}{b\ x\ c}$

if it's > 1 → there is risk

if it's < 1 → it's a protective factor

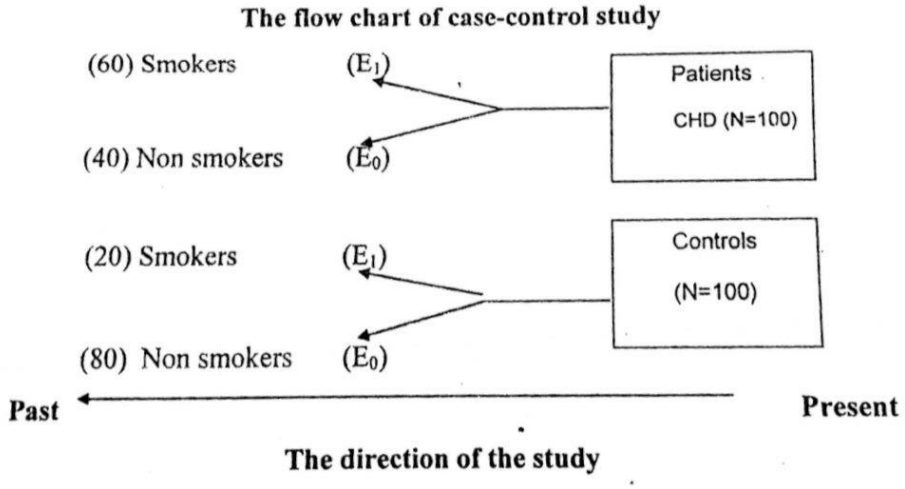
if it's = 1 → there is no relation between the factor & disease

Advantages of cross-sectional study:

1. Easy, quick, cheap
2. suitable for rare diseases
3. suitable for diseases with long latency period
4. you can check more than one exposure "risk factor" at the same study

Disadvantages:

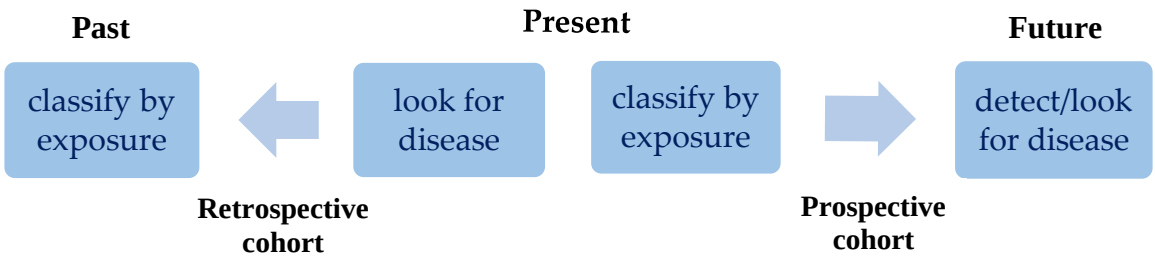
1. Selection bias
2. Recall bias → لأنها بتعتمد على الذاكرة فمش كل العيانين هيقدروا يفتكروا من امتى بدأوا يشربوا سجاير مثلاً
3. Not suitable for rare exposure

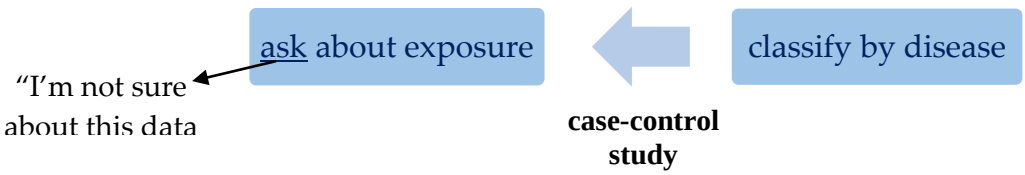


---

b.2. Cohort study:

- **Prospective**
- **Retrospective** (=case-control ولكن مع بعض الاختلافات)





Prospective Cohort study:

Selection of individuals on basis of being exposed or non-exposed and follow them for a certain period of time to determine whether or not the disease developed.

Steps of Cohort study:

1. Selection of cases:

a- Selection of exposed:

If the exposure is:

- Common (smoking – obesity …) → you select from general population
- Specific (rare) (uranium …) → you select from subgroups or certain occupations

b- Selection of non-exposed:

If the exposure is:

- Common → the control will be from the same population, & the comparison between exposed & non-exposed = **internal comparison**
- Specific to certain group → the control will be from another group of people & the comparison between exposed & non-exposed = **external comparison**

2. Obtain data on exposure:

a. from subjects
b. medical records
c. medical examination
d. lab investigations
e. environmental surveys

3. Follow up the 2 Cohort for the whole latency period:

4. Obtain data on outcome:

a. death records
b. medical records
c. medical examination
d. lab investigations

5. Analysis and interpretation:

Exposure	Disease	Healthy	Total
Yes	a	b	a + b
No	c	d	c + d

1. The incidence rate among exposed $I_e = \frac{a}{a+b}$ x constant
2. The incidence rate among non-exposed $I_0 = \frac{c}{c+d}$ x constant
3. Relative risk (RR) = $\frac{\text{incidence among exposed}}{\text{incidence among non-exposed}}$

   → *It measures the strength of association between suspected cause & effect*

4. Interpretation:

   if it's > 1 → the exposure is a risk

   if it's < 1 → it's a protective factor

   if it's = 1 → the exposure is not associated with the disease

5. The attributable risk percent (ARP) = $\frac{Ie - Io}{Ie}$ x 100

   → *It indicates to what extent the disease is attributed to exposure*

Retrospective Cohort study:

Here the investigator will go back years to select the study subject (Exposed **E_e**) and the comparison group (non-exposed **E_0**). Retrospective study is useful for occupational exposures.

Advantages of Cohort studies:

1. suitable for rare exposures
2. can check multiple outcomes of one exposure (smoking causes → lung cancer, CHD,…)
3. can calculate incidence of disease among exposed and non-exposed
4. can estimate relative & attributable risk
5. Dose response ratio can be calculated
6. No selection bias
7. Allows testing hypothesis & Temporal consequence between the exposure and disease can be proved (no chicken egg dilemma)

Disadvantages:

1. Not suitable for rare diseases

Prospective type

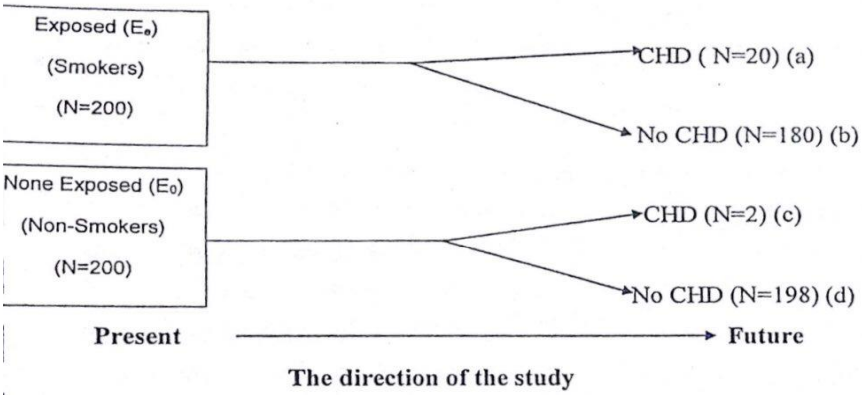
2. loss of staff & funds
3. change in environmental factors
4. change in diagnostic methods/criteria
5. study may alter participant behavior
6. loss of participants
7. ethical problems
8. expensive & time consuming

Retrospective type

9. records aren't available
10. Not accurate

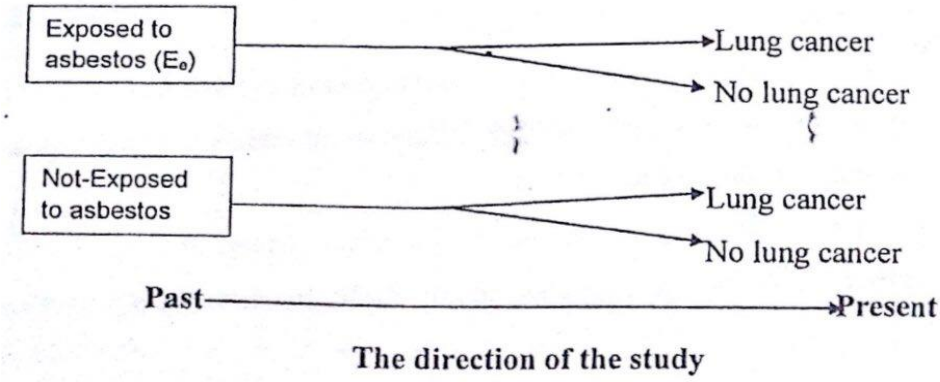
*Tables P.58 & P.60 → just for revision

The flow chart of prospective cohort study



The direction of the study

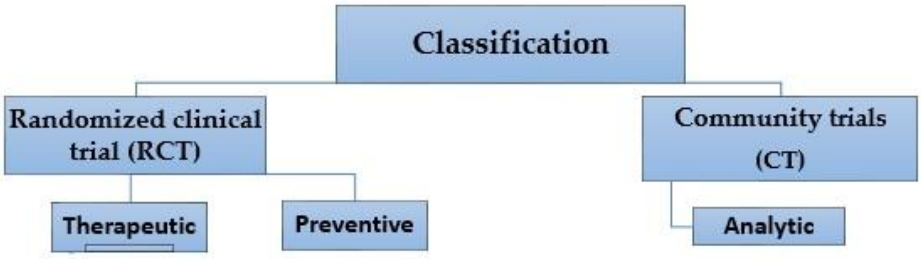
Flow chart of retrospective cohort study



The direction of the study

II- Intervention (experimental) Studies:

- It's a follow up study in which the intervention (exposure) is applied by the investigator
- It's the best epidemiological study to prove causality & effectiveness of a specific intervention



A- Randomized clinical trial (RCT):

- we have 2 groups:
  - experimental group = treatment group → *"who receive the drug or intervention"*
  - control group → *"who doesn't receive the drug or intervention"*

1. Therapeutic trial:

- conducted among patients with a particular disease
- To determine the ability of the intervention to:
  * reduce symptoms * prevent recurrence
  * decrease the risk of death from the disease

2. Preventive trial:

- conducted among free healthy persons
- To evaluate whether the intervention reduces the risk of developing a disease

Principles of clinical trials:

1. Control of variables ➔ this study is free of bias

As the investigator has the control to select experimental group, controls, interventions and conditions

2. Usage of comparison group:

- They are mainly for comparing the effect of the intervention on the experimental group in relation to them.
- They receive placebos (blank doses – saline injection – sugar pill).

3. Randomization:

* Assigning subjects to the experimental group in a completely random manner

→ To avoid the potential bias for the researcher or subjects

4. Blindness:

- the researcher is unaware of the identities of the experimental groups
- Also the subjects may be unaware if they are receiving treatment or placebo
- If neither researcher nor subjects know who is receiving the treatment this is called **Double blinded trial.** This is to avoid the possible bias

So the standard gold design for clinical trials "the least prone to bias" is
➔ Randomized double blinded controlled trials.

When we analyze data, it is done like Cohort analysis except:

Instead of calculating attributable risk, we can calculate "**Absolute risk reduction**"

	Outcome +	**Outcome -**
Intervention +	a	b
Intervention -	c	d

Absolute risk reduction =

Risk rate of developing outcome in intervention group - Risk rate of developing outcome in control group

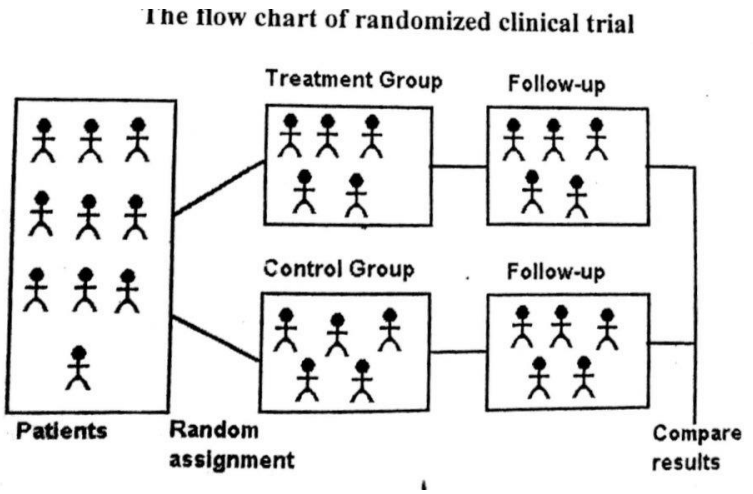
$= \left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right)$

Advantages:

1. Provide the strongest evidence of causality as they:
   a- ensure temporal relationship
   b- control for confounding through randomization
2. The best design to determine the effectiveness of a vaccine, drug or surgical procedure
3. The least prone to biases

Disadvantages:

1. Expensive & time consuming
2. Randomization may not be accurate when we study risk factors or prognostic factors
3. Ethical issues
4. Non-compliance
5. large sample size is needed



B- Community trials (CT):

Unit of the study is community (not individuals)

Examples:

- applying vaccines
- iron-fortified salts introduced into food
- information campaigns

Biostatistics

Statistics: the science that deals with collection, classification, presentation, description, analysis and interpretation of data.

Descriptive statistics:
- concerned with the summary measures of data collected from a sample of population

Analytic statistics:
- concerned with the use of data collected from a sample of population to make interference about population

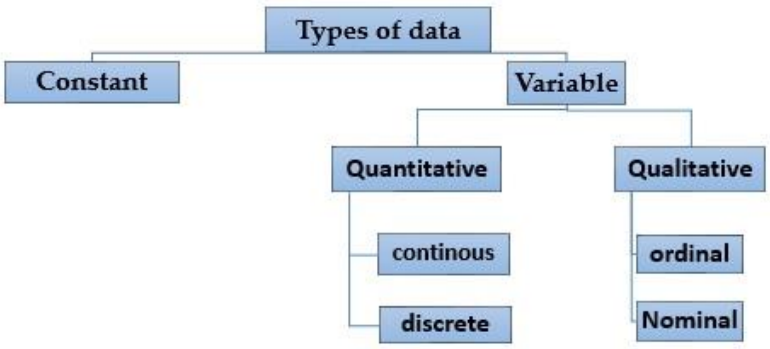
Vital statistics:
- ongoing collection by government agencies of data relating to events such as births, deaths, marriage…

Biostatistics: application of statistical procedure in the field of biology/medicine.

Uses of biostatistics in medicine:
1. To evaluate and criticize the researches published in medical journals
2. To assess diagnostic testing, effect of new drugs and treatment modalities
3. Epidemiologists need to calculate rates to compare between groups to decide if the difference between groups is due to chance

Data: Basic building blocks of statics.

Types of data:



1. **Constant data:**
   Observations don't vary from person to another, e.g. number of eyes/fingers…
2. **Variable data:**
   Observations vary from person to another
   a. **Quantitative data:** → something deals with numbers, Between each variable and the other there is a constant distance. يعني مثلاً الفرق ما بين الطول 164 سم و 165 سم ثابت = 10 مللي ، وبين 165 سم و 166 سم = 10 مللي وهكذا..
      1. Continuous:
         Obtained by **measurement**. Its value could be **integer** "عدد صحيح"or **fractional**"كسر"
         e.g. height – weight – age – Hb – income.

2. Discrete:
   Obtained by **enumeration**. Its value should be **integer** "عدد صحيح"
   e.g. pulse – family size – number of live births – abortion.

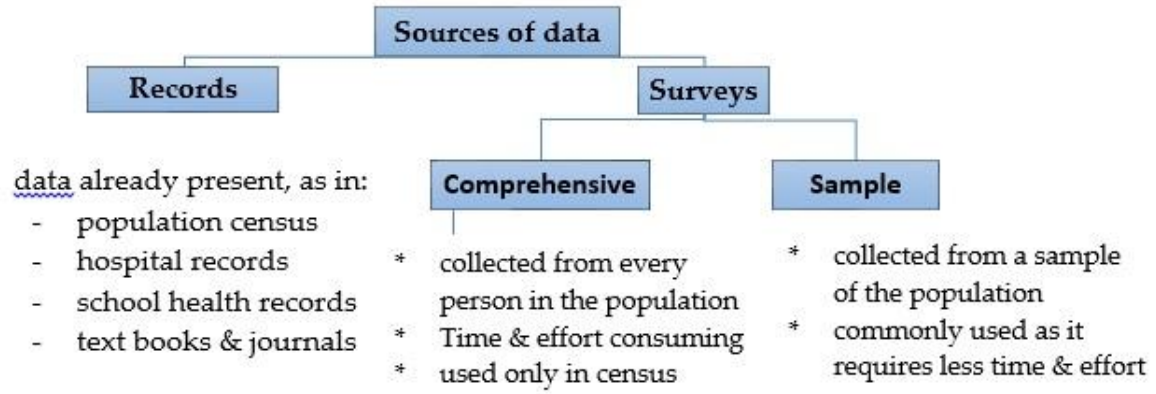
b. Qualitative data: → can't be measured or enumerated. Between each variable and the other, there is no constant distance.

تحديداً تظهر في جزئية ال scoring, ranking ،، يعني الفرق بين الأول على السباق والثاني ممكن يكون 3 متر أو 4 أو أي رقم كان، وكذلك الفرق بين التاني والتالت مش شرط يبقى زي الفرق بي الأول والتاني وهكذا..

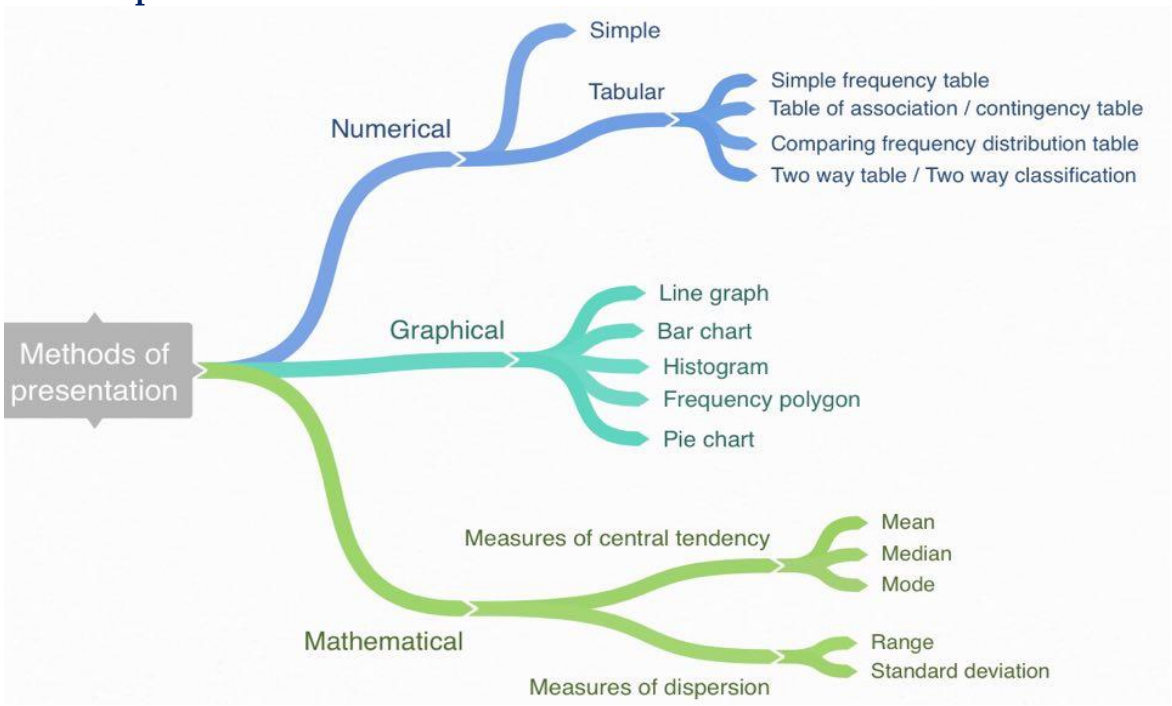
1. Ordinal:
   Can be put in order.
   e.g. degree of success

2. Nominal:
   Can't be put in order, and may be:
   - **Dichotomous**
     → sex – yes/no variables
   - **Multichotomous**
     → blood groups – marital status

- **How we collect data?**
- **Methods of collecting data:**
- **Sources of data:**



- **How we present our data?**



I- Numerical presentation:

a- Simple: (ungrouped / unclassified data)

This method is used when dealing with small size (5 – 7 – 10 observations)

e.g. weight of 5 infants – height if 7 children ….

b- Tabular: (grouped data)

The most convenient method for summarization of a large mass of data

1. Simple frequency table:

The variable may be:

Qualitative:

- you have to write the categories of the variables
- Frequency of observations of each category.

Degree of success	Number of students
Excellent	**120**
Very good	**200**
Good	**350**
Fair	**180**
Total	**850**

* دايماً ال variable بيتكتب في العامود اللي على الشمال .

Quantitative:

How to express your variable?

- <u>usually in the form of intervals:</u>

1- Number of intervals (4- 12)

2- Width or size of each interval usually (5 or 10 or 15)

3- The first interval begins with the smallest observation. (lower limit of 1st interval)

4- **We shouldn't let the table with open end:**
- lower limit of 1st interval should be known
- upper limit of last interval should be known

5- Notice that:
* when we use lower number of intervals e.g. 4 → we lose precision
* when we use higher number of intervals e.g. 12 → we lose summarization

6- Each interval has:
- lower limit
- upper limit
- width
- mid-point
- number of observations

- We can avoid duplication of limits of intervals through these methods:

"يعني منكررش قيمة خانتين واحنا مش واخدين بالنا"

A	B	C	D
10 to less than 15	10-14	10-	10-14.9
15 to less than 20	15-19	15-	15-19.9
20 to less than 25	20-24	20-	20-24.9
25 to less than 30	25-29	25<30	25-29.9

* A , C → can be used in both continuous & discrete data
* B → used only in discrete data
* D → used only in continuous data

Examples for sample frequency distribution table:

For **continuous quantitative:**

Hb%	No. of students
65-	22
70-	15
75-	19
80-	11
85-	17
90-	12
95<100	4
Total	**100**

First interval:

Lower limit = 65

upper limit = 69.999 ≈ 70

width = U.L. – L.L. = 70 – 65 = 5

Mid-point = $\frac{UL+LL}{2} = \frac{70+65}{2} = 67.5$

Number of observations = 22

For **continuous quantitative:**

Family size	No. of patients
3 - 4	28
5 – 6	55
7 – 8	32
9 – 10	18
11 - 12	12
Total	**145**

Last interval:

Lower limit = 11

upper limit = 12

width = **2**

هنا بتعد الitems في كل interval بنفسك ،، يعني لو كانت 10-14 ← 5

Mid-point = $\frac{UL+LL}{2} = \frac{11+12}{2} = 11.5$

Number of observations = 12

---

2. **Table of association / contingency table:**
   Used to show relation between condition & characteristic (risk factor)
   "Causality table"
   it may be : Two x Two **or** c x r "more than two"
3. **comparing frequency distribution table:**
   When comparing one variable in 2 groups
4. **Two way table or two way classification:**
   When we find correlation between 2 variables in
   * table P.80

اسمه كده عشان هنا عندي 2 variables بكتب واحد في عمود على الشمال والتاني في صف عرضي، والobservation بتكون طبقاً لحاجتين مش حاجة واحدة

II- Graphical presentation:

1. **Line graph** → can be used when the variable is **Time** (which is a special form of continuous quantitative variables)
2. **Bar graph** → used for:
   - qualitative variables (ordinal – nominal)
   - quantitative variables (discrete)
3. **Histogram** → used for:
   - quantitative continuous variables "*the table should be of **simple frequency distribution** type*"

4. **Frequency polygon** → used for:
   - quantitative continuous variables "*the table may be of **simple frequency distribution** type or **any other complex** tyoes*"

5. **Pie chart**→ can be used for **<u>all types of variables</u>** ,
   but it isn't the best to present any of them

Notes:

- We have 2 axes (X & Y)
  → always the variable (which in the left column) is presented on the X axis

1. **Line graph:**
   - We put a point for each exact time value and the corresponding value on the vertical axis.
   - Then every 2 consecutive points are joined by straight line.

   e.g. - Temperature chart of patients
   - changes in trends of birth and death rates

2. **Bar chart:**
   - Each category of the variable is represented by a bar, its height is corresponding to the value on the vertical axis
   - In case of complex tables → we may draw 2 bars for each category of variables

3. **Histogram:**
   - Each interval is represented by a column, its 2 vertical line represent lower limit and upper limit of the interval, and its height equals the value on vertical axis
   - There is no spaces between the consecutive columns.

4. **Frequency Polygon:**
   - Each interval is represented by a point which is " mid point" and the corresponding value in the vertical axis.
   - Then every 2 consecutive points are joined by a straight line.
   - If we have a complex table type , then we will draw 2 lines or more.
   - There is no spaces between the consecutive columns.

5. **Pie chart:** " كنا بنرسمها فى ابتدائى^_^ "
   - The circle is divided into sectors , the number of sectors is equal to number of categories or intervals.
   - Each sector has a percent of the circle.
   - Each sector has an angle which is calculated from dividing the frequency of this category or interval over total frequency then multiply the product to 360 .
   - We start from 12 o'clock and go in a clockwise direction.

III- Mathematical presentation:

a- Measures of central tendency:

computed values around which most of the observations are collected.

	Arithmetic mean	Median	Mode
1-Symbols	$\bar{X}$ = sample mean μ = population mean	~ X = median	*Mo* (Salah :D)
2- How to calculate	In ungrouped data: e.g: 2,3,4,5 $\bar{X}$ = sum of observations / its number $= \frac{\sum X}{n} = \frac{2+3+4+5}{4} = \frac{14}{4}$ = 3,5 In grouped data: In case of quantitative continuous data: You have to calculate midpoint of the interval (X) $\bar{X} = \frac{\sum(fX)}{\sum F}$ P.90	In ungrouped data: When number of observations is odd: We arrange observations in a descending or ascending manner , then get the middle observation. Rank of middle value= $\frac{n+1}{2}$ e.g: 2,3,6,8,9 Third value , ~X=6 When number of observations is even: We arrange the values …. then the 2 middle values are summed and divided by 2 . Rank of the 2 middle values = $\frac{n}{2}$, $\frac{n+1}{2}$ e.g: 2,3,4,5,6,10 third and fourth values ~X = $\frac{4+5}{2}$ = 4,5	In ungrouped data: The most frequent observations , e.g: 2,2,2,3,4,5 Mode = 2 In grouped data: the most frequent observation : In quantitative continuous data ➔ mid point of the interval with highest frequency. In quantitative discrete or qualitative data ➔ the category with highest frequency.

3- Advantages	1-takes all observations into consideration. 2-best average to be used in statistical analysis.	1-can be used with quantitative and qualitative ordinal variables. 2-can be used in open ended table. 3- not affected by extreme observation.	1-can be used with all types of variables. 2- can be used in open ended table. 3-Not affected by extreme observations.
4. Disadvantages	1-can't be used with qualitative variables. 2-easily affected by extreme observations. 3-can't be used with open ended table.	1-can't be used with qualitative nominal variable. 2-not easy to be used in statistical analysis.	1-It may not be able to be get if observations have equal frequencies. 2-sometimes we may get 2 modes or multiple modes. 3-not easy to be used in statistical analysis.

b- Measures of Dispersion:

- They measure how far the observations from the mean.

1. Range:

It is difference between the biggest and smallest observations.
= it is difference between the upper limit of last interval and lower limit of first interval.

Examples:

- 30, 34, 32 , 36 and 28 years
  Mean = 32
  Range = 36-28 = 8

- 12, 30 , 8 , 62 . 42 years
  Mean = 32
  Range = 62-8= 54

Although the 2 groups have the same mean , but it doesn't express the mean of the second group in reality.

2. Standard of deviation (S) ,(σ_x) , (z):

The best commonly used measure of dispersion .
It is the positive square root of the variance (S^2).
It measures the deviation of observations from the mean.
measurements of standard of deviation in ungrouped data and grouped data "page 96, 97 and 98" .

Normal Distribution Curve
(Gussian distribution curve)

Characteristics of the curve:

1- Bell shaped curve.

2- Has lower and upper tails.

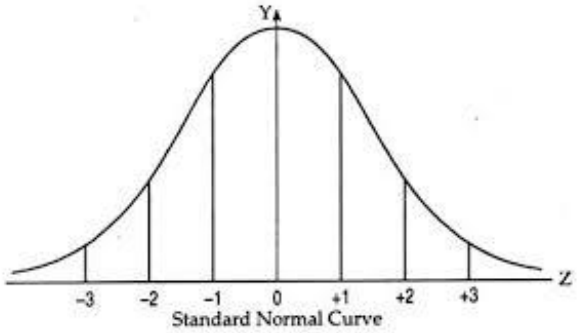
3- Mean = Median = Mode , all lie in the centre of the curve.

4- 68% of population lie between +/- 1 SD (standard of deviation) from the mean .
28% are distributed as : 14% between +/- 1SD and 2 SD from the mean

- We can say that nearly 95% lies between +/- 2 SD from the mean.
- 99% lies between +/- 3 SD from the mean.
- Many human traits are distributed among population in normal way.
- If we want to calculate the confidence interval ➔ which is the area that include 95% or 99% of population you can calculate it as follows:
- If you want to know the limits where 95% of population lies on the curve :
  Mean +/- 1,96 z (z is the standard of deviation)

- If you want to know the limits where 99% of population lies on the curve :
  Mean +/- 2,58 z (z is the standard of deviation)

طب ليه بنضرب فى 1.96 أو 2.58 دى ليها معادلات كتير و حسابات طويلة ولكن يكفى بس اننا نبقى عارفين ان الحدود دى اللى هيبقى فيها 95 % من الناس أو 99% حسب انا حسبت انهى



Standard Normal Curve

Statistical Analysis

As we said before there are types of statistics :

1- Descriptive ➔ concerned with summarization and presentation of data.

2- Analytical ➔ concerned with usage of data to make a decision.

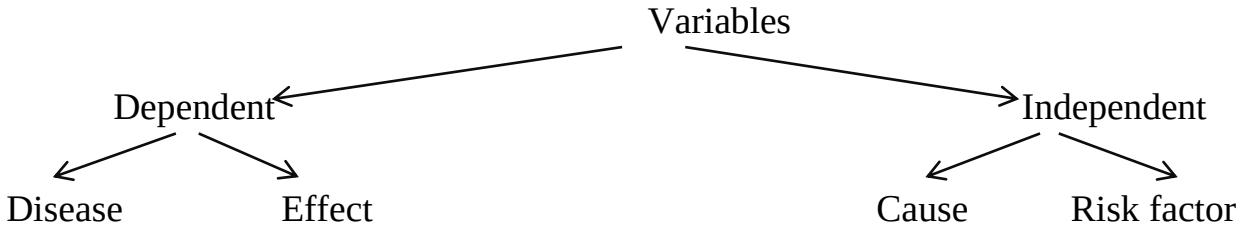
Types of statistical analysis:

a) Differences: compare one variable (dependent) = Disease or effect .

يعنى هنا بيقارن مثلا تأثير الدوا أ و الدوا ب ، و هل فى اختلاف فى التأثير بتاعهم على نفس المرض مثلا أو بيقارن نسبة الهيموجلوبين فى مجموعه معينه مقارنة بالنسبة المتعارف عليها بين الناس

b) Associations: compare 2 or more variables (at least one of them must be dependent and others independent within the same group.

يعنى هنا مثلاً بنشوف هل فى ارتباط بين استخدام القهوة كتير (اندبندت) و ارتفاع ضغط الدم (دبندنت) ولا لأ ؟



Steps of statistical analysis:

1- Formulate hypothesis:

- Null hypothesis (H_0)➔ no significant difference / association.

In statistical analysis we test mainly the null hypothesis.

2- Determine the type of statistical type.

3- Determine the significance.

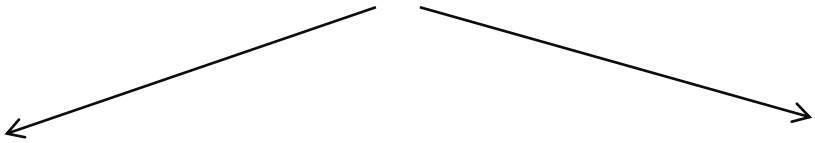
How to determine type of statistical test :

a- Type of dependent variable , distribution of the quantitative one ➔ (normal – abnormal)

We use parametric tests — we use non – parametric tests.

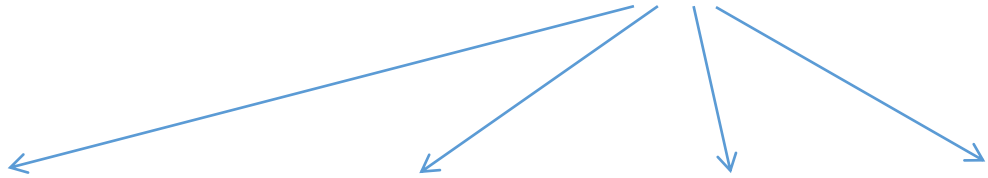
b- Numbers of groups.

Statistical tests of differences



Qualitative dependent variables (presented as N or %)
Chi square test (X^2 test)

Quantitative dependent variables
(presented as mean and it is normally distributed) t – test



one group population	one group (before and after)	2 different groups	>2groups
t- test population	paired t test	student's t test	f -test

في الامتحان مش هيبقى مطلوب منا نحفظ القوانين او نحسب الt value

Level of significance (α) :

- It indicates the rejection area in the hypothesis.
- You determine it before starting the study.
- Usually , it is 5% or 1%.

1- Degree of freedom :

In case of one group ➔ DF= n-1
2 groups ➔ n_1+n_2-2
table formed of c*r cells ➔ (c-1)*(r-1)

2- Critical Value:

Computed value depends on the level of significance and degree of freedom.

- We compare the value of the test to the critical value :

If >= critical value ➔ reject the null hypothesis
If < critical value ➔ accept null hypothesis.

3- P- value (probability of chance):

If P. value < 5% ➔ we reject the null hypothesis.
If > = 5% ➔ we accept the null hypothesis.

أى مسألة تيجى مطلوب منى أعمل فيها ايه ؟
أول حاجة أحدد نوع ال تست هو انهى واحد من ال 5
أكتب ال فرضيتين بتوعى

Null hypothesis : there is no statistical significant difference between ……

Alternative hypothesis: there is statistical significant difference between ……

أقدر أخد قرار وهو إنى أقبل أو أرفض الفرضيتين عن طريق مقارنة ال(t value) اللى هيديهانى بال critical value اللى هيديهالى برضو .

you need to know about chi square that:

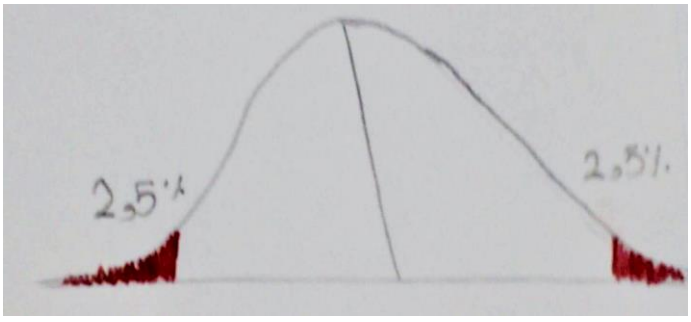
- The expected values should be 5 or more.
- The total number or grand total should be more than 30.

كل اللي فات احنا بنجمع بس هنعمل ايه مع كل نوع من البيانات اللي عندنا ونستخدم معاها انهي test،، دلوقتي لازم نفهم كام exception كده عشان اللي بعد كده يبقى سهل ونطبقه عن فهم.
^^ "ll summarize it anyway't worry I't interested you can skip this part and don'If you aren "

1. Level of significance:

هنا في الstatical test انا بختبر الnull hypothesis ، وبالتالي لو انا اخدت sample from the population وقولت ان النتايج اللي طلعتها دي عايز أطبقها على الpopulation كلها ولكن في برضو نسبة من الحظ ممكن تطلع وتغلط النتايج دي وبالتالي هنرفض الnull hypothesis

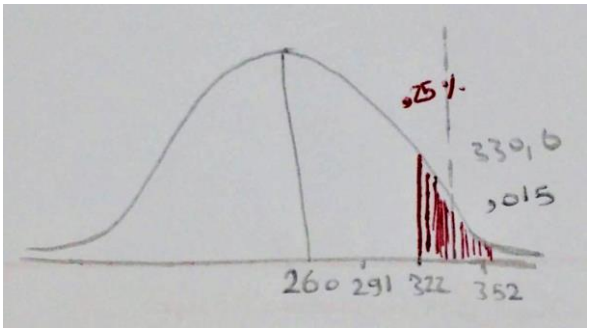
- في الرسم التالي لو اعتبرنا ان Ho صحيحة وبرضو لازم نحط في اعتبارنا ان ممكن تطلع نسبة بالحظ كده فيها اختلاف، فمش هنقول بقى اننا نرفض النظرية بحالها وأنها غلط، لأ احنا هنقول ان عندنا فعلا Ho صحيحة وفِي نسبة معينة كده (level of significance) ممكن يحصل فيها اختلاف وساعتها مش هينطبق عليها الHo ولكنه هيكون مقبول على كل حال لأن انا اللي يحدده من قبل ما أفرض النظرية أصلاً.



- if we set a level of significance (α) = 5% : that means our hypothesis can be applied on 95% of population, while 5% can't apply this hypothesis (reject the hypothesis)
- Level of significance (α) is mainly 5% or 1%

2. P. Value (probability of chance):

- دلوقتي انا حطيت Ho بتاعتي وقلت مثلا ان مفيش اختلاف بين معدل الذكاء في الpopulation ومعدل الذكاء في مجموعة من طلاب كلية الطب مثلاً.
- رسمت normal distribution وطلع عندي الmean مثلا 260 ، ولما قسمت المعدل لمجموعة الطلاب طلع الmean عندي 330.6.
- لو انا بقى عايز احسب الP. Value بتاعت مجموعة الطلاب



- أيوه يعني ايه P. Value ← يعني المجموعة دي تمثل فعلياً كام % من الpopulation على ال normal distribution curve
- الحسابات معقدة جدا الحقيقة ولكنها ممكنة ومش موضوعنا على كل حال
- هنا ال P. Valueهتساوي في الطرف الواحد حوالي 0.01556. (%1.5) يعني في الناحيتين حوالي 0.03112 . (%3)
- هنا بقى انا هقارن بين الP. Value والα اللي انا اتفرضتها من البداية خالص:

- لو انا كنت فرضتها 5% —> يبقى كده المجموعة دي فعلا وقعت في المساحة بتاعت الα وهرفض الHo
- أما لو كنت فرضتها 1% —> يبقى المجموعة دي عادي جداً لسه في الconfidence interval وهقبل الHo

3. Degree of freedom:

- هناخد مثال سريع يوضح لنا المفهوم ده

- لو انا قلتلك ان إشارة المرور لونها مش احمر ولا اصفر ، تفتكر هتبقى ايه؟ !!😂 أكيد أخضر
- هنا انا اعتمدت على معلومتين عشان اجيب التالتة
- ولو المعلومتين دول اتغيروا هيأثروا بالتأكيد على التالتة

D.F. : number of independent pieces of information on which the parameter estimate is based.

- DF for one group = n - 1
- DF for 2 groups = $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$
- DF for (2x2) table = (c – 1) (r – 1)

4. Value of the test: (t)

دي قيمة انا بحسبها حسب نوع الtest اللي شغالين بيه وبستفيد منها بعد كده في مقارنتها بحاجة اسمها critical value

5. Critical value:

دي قيمة محسوبة مسبقاً ويطلعها من جدول معمول جاهز، بتعتمد على معرفتي لحاجتين هما الα والDF

- لو الt طلعت اكبر من او تساوي الcritical value فكده we will reject the Ho
- أما لو طلعت أقل منها فكده we will accept the Ho

نستنتج من كل ده بقى ان لو عايز أ make a decision في الstudy بتاعتي قدامي حل من اتنين:

١- يا اما احسب الt value وأقارنها بال critical value

٢- أو يبقى عندي الP. Value وأقارنها بالα